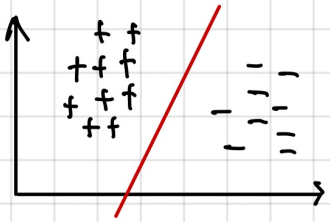
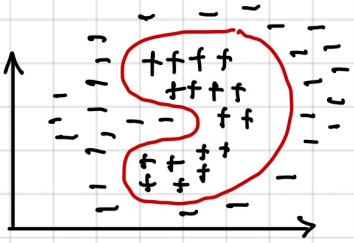
LECTURE 5 : CLASSIFICATION

Motivation : Besides regression, classification is another important factor in making decisions. In fact, most of decisions involve either regression or classification.

Examples of classification :





[1] Classification Linear Model

Data modeling :

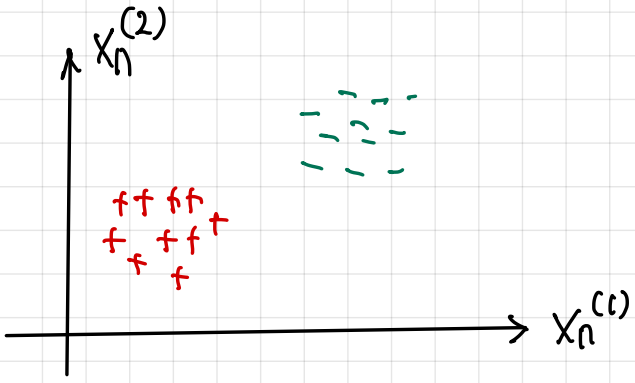
$$y(\bar{x}_n) = \sum_{d=0}^{D} w_d x_n^{(d)} = w_0 + \sum_{d=1}^{D} x_n^{(d)}$$

$$= \bar{w}^T \bar{x}_n$$

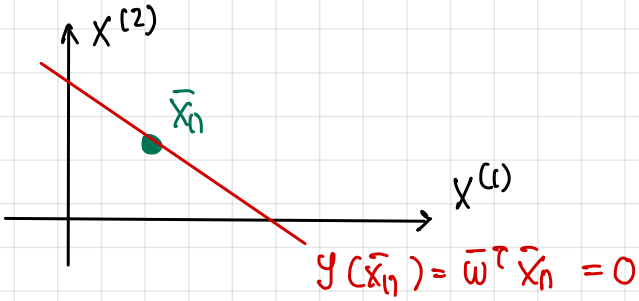
where : $x_n^{(0)} = 1$

$$\bar{x}_n = \begin{bmatrix} 1 \\ x_n^{(1)} \\ x_n^{(2)} \\ \vdots \\ x_n^{(D)} \end{bmatrix}$$

; D = Dimensionality of the classification space



Consider a sample $\bar{x}_n$: $y(\bar{x}_n) = \bar{w}^T \bar{x}_n = 0$



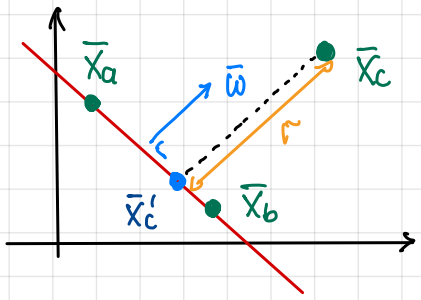
$$\bar{x}_n = \begin{bmatrix} 1 \\ x_n^{(1)} \\ x_n^{(2)} \end{bmatrix}$$

$$y(\bar{x}_n) = \bar{w}^T \bar{x}_n = \begin{bmatrix} w_0 & w_1 & w_2 \end{bmatrix} \begin{bmatrix} 1 \\ x_n^{(1)} \\ x_n^{(2)} \end{bmatrix} = 0$$

$$= w_0 + w_1 x_n^{(1)} + x_n^{(2)} = 0$$

Recall: $ax + by + c = 0 \longrightarrow y = -\frac{a}{b}x - \frac{c}{b}$: Line equation

Consider two samples $\bar{x}_a$ and $\bar{x}_b$:



$$y(\bar{x}_a) = y(\bar{x}_b) = 0$$
$$y(\bar{x}_a) - y(\bar{x}_b) = 0$$
$$\bar{w}^T(\bar{x}_a - \bar{x}_b) = 0$$

Recall: $\bar{a}^T\bar{b} = \bar{a} \cdot \bar{b} = \|a\|\|b\| \cos\theta$

if $\bar{a}^T\bar{b} = 0$, then $\theta = 90°$

$(\bar{x}_a - \bar{x}_b)$ is a line segment lying on the red line, $y(\bar{x}_a) = 0$.

Hence: $\bar{w}$ is perpendicular to the red line.

Assume $\bar{x}_c$, which is not on the line : $y(\bar{x}_c) \neq 0$

$$\bar{x}_c = \bar{x}_c' + r \, \bar{w}/\|\bar{w}\|$$

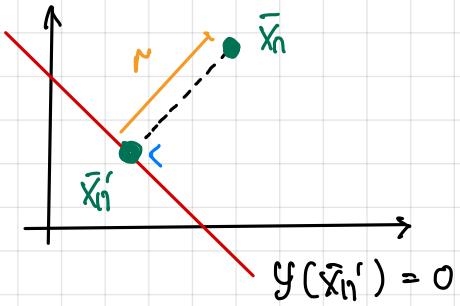
$$y(\bar{x}_c) = \bar{w}^T \left(\bar{x}_c' + r \frac{\bar{w}}{\|\bar{w}\|} \right)$$

$$= \underbrace{\bar{w}^T \bar{x}_c'}_{=0} + \bar{w}^T r \frac{\bar{w}}{\|\bar{w}\|}$$

$$y(\bar{x}_c) = r \frac{\bar{w}^T \bar{w}}{\|\bar{w}\|} = r \frac{\|\bar{w}\|^2}{\|\bar{w}\|} = r\|\bar{w}\|$$

($\bar{w}^T\bar{w} = \|\bar{w}\|\|\bar{w}\|\cos 0 = \|w\|^2$)

If $y(\bar{x}_n) = 1$, then:



$$y(\bar{x}_n) = \bar{w}^T \bar{x}_n = r \|\bar{w}\| = 1$$

$$r = 1 / \|\bar{w}\|$$

$\Downarrow$

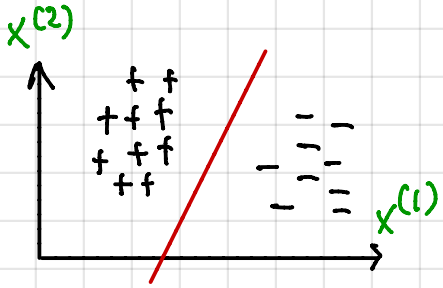
$y(\bar{x}_n) = 1$ implies $\bar{x}_n$ is located outside the line with distance r.

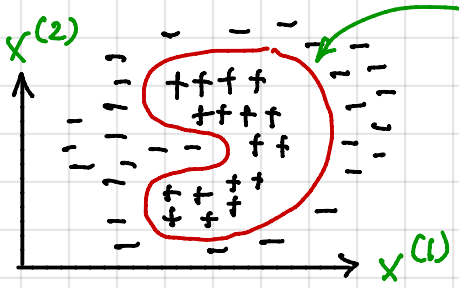
SPACE TRANSFORMATION

So far, we define: $\underset{K\times 1}{\bar{y}(\bar{x}_n)} = \underset{K\times D}{W^T}\ \underset{D\times 1}{\bar{x}_n}$.

This comes with a drawback:

Our boundary line is always linear, however the distributions of the data belonging to class 1 & 2 might not be linear:





This can't be solved in the $\bar{x}$ space

To address this problem, we need to transform $\bar{x}_n$ into another (usually higher) space:

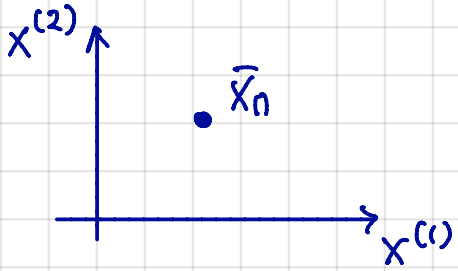
$$\underset{1\times 1}{y_k(\bar{x}_n)} = \underset{1\times M}{\bar{w}_k^T}\ \underset{M\times 1}{\bar{\phi}(\bar{x}_n)} = \sum_{m=0}^{M} w_k^{(m)}\,\phi^{(m)}(\bar{x}_n)$$

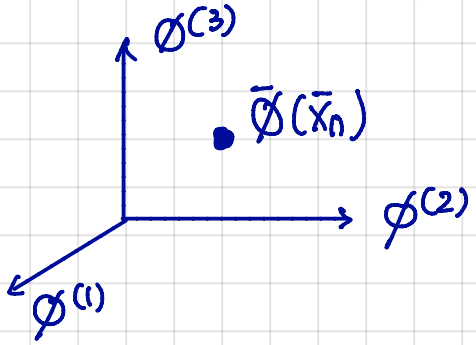
e.g.: $\phi^{(m)}(\bar{x}_n) = \exp\left(\frac{-\|\bar{x}_n - \bar{\mu}_m\|_2^2}{2s^2}\right)$

$\bar{a} = \begin{bmatrix} a_1 \\ a_2 \end{bmatrix}$; $\|\bar{a}\|_2^2 = \left[\sqrt{a_1^2 + a_2^2}\right]^2$

$\bar{\mu}_m$ is a vector with the same elements/values:

$\|\bar{x}_n - \bar{\mu}_m\|_2^2 = (x_n^{(1)} - \mu_m)^2 + (x_n^{(2)} - \mu_m)^2$; $\bar{\mu}_m$ is the centers of clusters





$$\underset{K\times 1}{\bar{y}(\bar{x}_n)} = \underset{K\times M}{W^T}\ \underset{M\times 1}{\bar{\phi}(\bar{x}_n)} \longrightarrow \underset{N\times K}{Y} = \underset{N\times M}{\Phi}\ \underset{M\times K}{W}$$

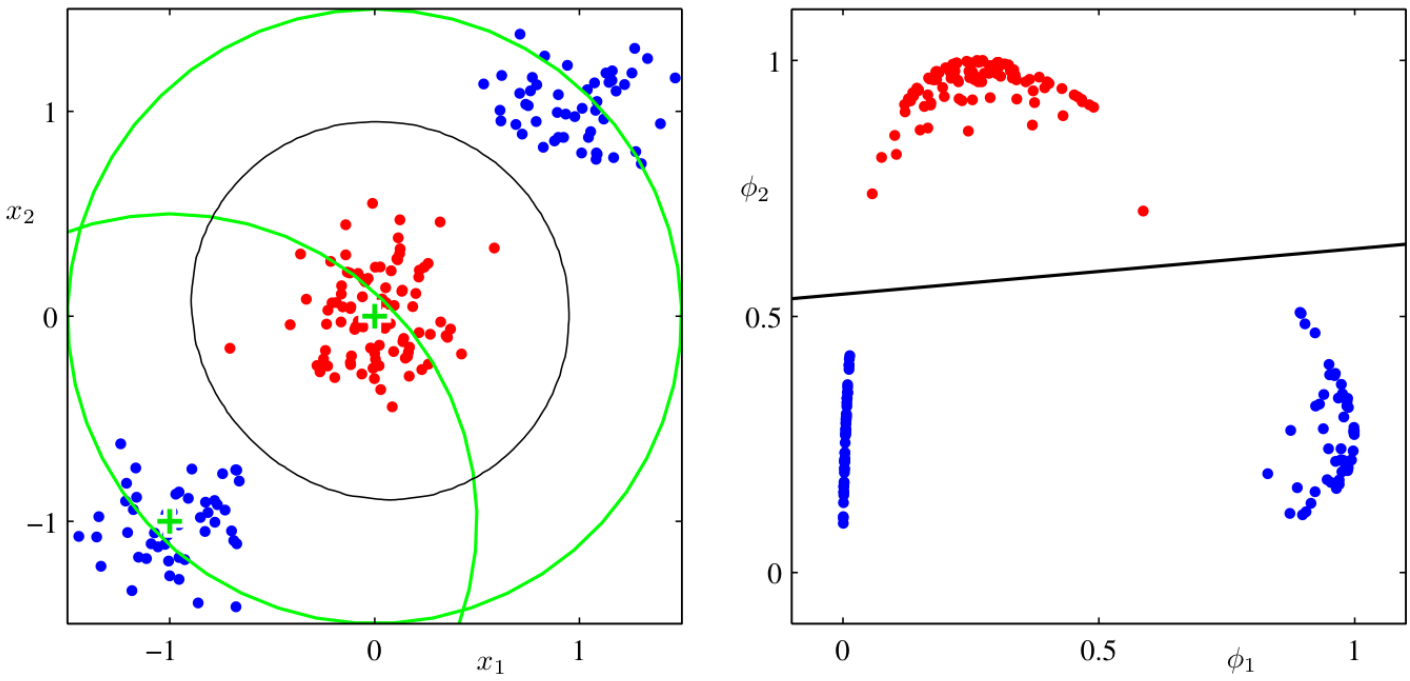


Figure 4.12 Illustration of the role of nonlinear basis functions in linear classification models. The left plot shows the original input space (x_1, x_2) together with data points from two classes labelled red and blue. Two 'Gaussian' basis functions $\phi_1(\mathbf{x})$ and $\phi_2(\mathbf{x})$ are defined in this space with centres shown by the green crosses and with contours shown by the green circles. The right-hand plot shows the corresponding feature space (ϕ_1, ϕ_2) together with the linear decision boundary obtained given by a logistic regression model of the form discussed in Section 4.3.2. This corresponds to a nonlinear decision boundary in the original input space, shown by the black curve in the left-hand plot.

MLE FOR CLASSIFICATION (LOGISTIC REGRESSION)

Testing stage: Input: $\bar{x}_*$ and $\bar{w}$

Output: t_* (label of $\bar{x}_*$) ; $t_* \in \{0, 1\}$

Assuming two classes only
0 = class 1
1 = class 2

Goal: $p(t_* | \bar{x}_*, \bar{w})$

(1) We want to know $p(t_* = 0 | \bar{x}_*, \bar{w})$ and $p(t_* = 1 | \bar{x}_*, \bar{w})$

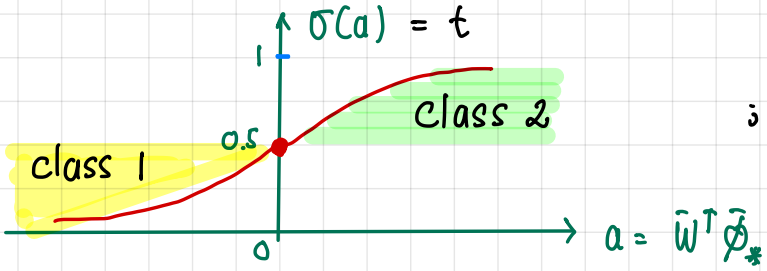
(2) if $p(t_* = 0 | \bar{x}_*, \bar{w}) \geq p(t_* = 1 | \bar{x}_*, \bar{w})$

then $t_* = 0 \rightarrow \bar{x}_*$ belongs to class 1

else $t_* = 1 \rightarrow \bar{x}_*$ belongs to class 2

Q: $p(t_* | \bar{x}_*, \bar{w}) = ?$ How can we model the probability function?

A: $p(t_* | \bar{x}_*, \bar{w}) = \sigma(\bar{w}^T \bar{\phi}(\bar{x}_*))$; $\sigma(a) = \frac{1}{1 + \exp(-a)}$

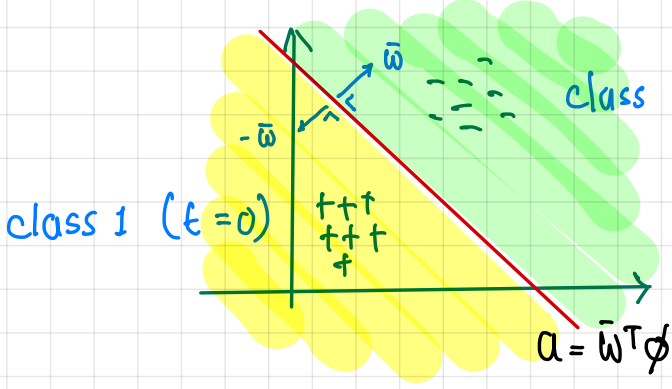


; $t = \{0, 1\}$

1 → class 1 ; ↳ class 2

$t_* = 0$: $a = w^T \phi_*$ is encouraged to be negative & large : $-a = -\bar{w}^T \bar{\phi}_*$

$t_* = 1$: $a = \bar{w}^T \phi_*$ is encouraged to be positive & large : $a = \bar{w}^T \bar{\phi}_*$



we only have one set of $\bar{w}$

Training stage: Input: $\{\bar{x}_n, t_n\}_{n=1}^{N}$

Output: $\bar{w}$

① Likelihood data modeling:

Likelihood: $p(t_n \mid \bar{w})$ → one sample

$p(t_n = 1 \mid \bar{w}) = \sigma(\bar{w}^T \bar{\phi}_n) = y_n$

$p(t_n = 0 \mid \bar{w}) = 1 - \sigma(\bar{w}^T \bar{\phi}_n) = 1 - (y_n)$

To make them into 1 equation:

$$p(t_n \mid \bar{w}) = y_n^{t_n} (1 - y_n)^{(1 - t_n)}$$

For all N data:

$$p(\underset{N\times 1}{\bar{t}} \mid \underset{M\times 1}{\bar{w}}) = \prod_n y_n^{t_n} (1 - y_n)^{(1 - t_n)}$$

Q: How do we know that it's a probability function?

A: For simplicity, let $N = 2$:

$p(t_1 = 1, t_2 = 1 \mid \bar{w}) = y_1 y_2$

$p(t_1 = 1, t_2 = 0 \mid \bar{w}) = y_1 (1 - y_2) = y_1 - y_1 y_2$

$p(t_1 = 0, t_2 = 1 \mid \bar{w}) = (1 - y_1) y_2 = y_2 - y_1 y_2$

$p(t_1 = 0, t_2 = 0 \mid \bar{w}) = (1 - y_1)(1 - y_2) = 1 - y_1 - y_2 + y_1 y_2$

+

$$\sum_{t_1} \sum_{t_2} p(t_1, t_2 \mid \bar{w}) = 1$$

② Error Function:

$$\bar{w}^* = \underset{\{\bar{w}\}}{\arg\max}\ p(\bar{t}\,|\,\bar{w}) = \underset{\{\bar{w}\}}{\arg\min}\ E(\bar{w})$$

$$E(\bar{w}) = -\log p(\bar{t}\,|\,\bar{w})$$
$$= -\sum_n \left(t_n \log y_n + (1-t_n)\log(1-y_n) \right)$$
$$= -\sum_n \left[t_n \log \sigma(\bar{w}^T\bar{\phi}_n) + (1-t_n)\log\left(1-\sigma(\bar{w}^T\bar{\phi}_n)\right) \right]$$

③ Minimization:

$$\frac{\partial E(\bar{w})}{\partial \bar{w}} = 0 = -\sum_n^N \frac{\partial}{\partial \bar{w}} \left[t_n \log \sigma + (1-t_n)\log(1-\sigma) \right]$$

Note: $\frac{\partial}{\partial x}\sigma(x) = \sigma(1-\sigma)$

$\frac{\partial}{\partial x}(1-\sigma(x)) = \sigma(\sigma-1)$

} see the last page for the proof.

$$= -\sum_n t_n \frac{\sigma(1-\sigma)}{\sigma}\bar{\phi}_n - \sum_n (1-t_n)\frac{\sigma(\sigma-1)}{1-\sigma}\bar{\phi}_n = 0$$

$$= \sum_n -t_n\bar{\phi}_n + t_n\sigma\bar{\phi}_n + \sigma\bar{\phi}_n - t_n\sigma\bar{\phi}_n = 0$$

$$\frac{\partial E(\bar{w})}{\partial \bar{w}} = \boxed{\nabla E(\bar{w}) = \sum_n (\sigma - t_n)\bar{\phi}_n} = 0$$

Q: How can we get $\bar{w}$ from the last equation?

A: Unfortunately, we cannot solve it in a closed-form manner.
Because $\sigma(-\bar{w}^T\bar{\phi}_n)$ is not linear.

[●] Iterative Reweighted Least Squares

Newton's method (or Newton-Raphson method):

$$\bar{w}^{new} = \bar{w}^{old} - H^{-1} \nabla E(\bar{w})$$

where: H is the Hessian matrix, which is $\frac{\partial^2 E(\bar{w})}{\partial \bar{w}^2}$ or $\frac{\partial}{\partial \bar{w}} \nabla E(\bar{w})$

$$\nabla E(\bar{w}) = \underset{M\times 1}{} \sum_n (y_n - t_n)\phi_n = \sum_n (\sigma(\bar{w}^T\phi_n) - t_n)\phi_n$$
$$= \underset{M\times N}{\Phi^T}\underset{N\times 1}{(\bar{y} - \bar{t})}$$

$$\underset{M\times M}{H} = \frac{\partial}{\partial \bar{w}} \left(\sum_n (\sigma_n - t_n)\bar{\phi}_n\right) = \sum_n \underset{M\times 1}{\bar{\phi}_n}\ \underset{1\times 1}{\sigma_n(1-\sigma_n)}\ \underset{1\times M}{\bar{\phi}_n^T}$$
$$= \underset{M\times N}{\Phi^T}\underset{N\times N}{R}\underset{N\times M}{\Phi}$$

R is a diagonal matrix

where: $R_{nn} = \sigma_n(1-\sigma_n)$

$$\bar{w}^{new} = \bar{w}^{old} - H^{-1}\nabla E(\bar{w})$$
$$= \bar{w}^{old} - (\Phi^T R \Phi)^{-1}\Phi^T(\bar{y} - \bar{t})$$
$$= (\Phi^T R \Phi)^{-1}\left[\Phi^T R \Phi\, \bar{w}^{\tau} - \Phi^T(\bar{y} - \bar{t})\right]$$
$$= (\Phi^T R \Phi)^{-1}\Phi^T R\, \bar{z} \quad \rightarrow \text{least-squares solution: } (\Phi^T\Phi)^{-1}\Phi^T\bar{t}$$

where: $\bar{z} = \Phi\bar{w}^{\tau} - R^{-1}(\bar{y} - \bar{t})$

Hence the name: Iterative reweighted least-squares

For our binary classification case we can initially set $\bar{w} = 0$

Note:

$$\frac{d\sigma(x)}{dx} = \frac{d}{dx}\left(\frac{1}{1+\exp(-x)}\right) = \frac{d}{dx}\left(1+\exp(-x)\right)^{-1}$$

$$= -\left(1+\exp(-x)\right)^{-2}(-\exp(-x)) = \frac{\exp(-x)}{(1+\exp(-x))^2} = \sigma^2\exp(-x)$$

$$= \sigma^2\left(1+\exp(-x)-1\right) = \sigma\left(\frac{1+\exp(-x)-1}{1+\exp(-x)}\right)$$

$$= \sigma\left(1-\frac{1}{1+\exp(-x)}\right)$$

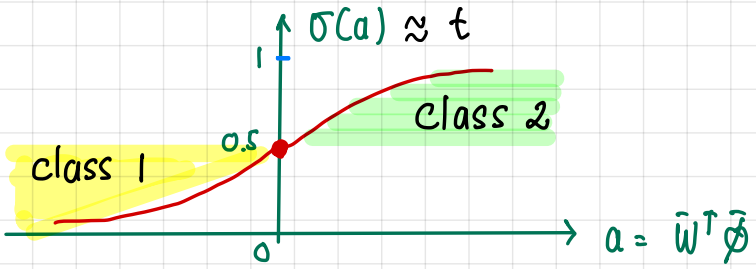
$$= \sigma(1-\sigma)$$

Accordingly: $\frac{d}{dx}\left((1-\sigma)\right) = -\sigma(1-\sigma) = \sigma(\sigma-1)$

MAP FOR CLASSIFICATION

Previously we know : $t_n \in \{0, 1\}$

$$t_n \approx y_n = \sigma(\underset{1\times M}{\bar{w}^T}\underset{M\times 1}{\bar{\phi}(\bar{x}_n)}) = \sigma(\underset{1\times M}{\bar{w}^T}\underset{M\times 1}{\bar{\phi}_n})$$



if $t_n = 0$: $a = \bar{w}^T\bar{\phi}_n$ is encouraged to be large & negative

if $t_n = 1$: $a = \bar{w}^T\bar{\phi}_n$ is encouraged to be large & positive

Unlike least squares, we have only 1 line, as we have only 1 set of $\bar{w}$.

- Training stage using MAP : Input : $\{\bar{x}_n, t_n\}_{n=1}^N$ Output : $\bar{w}$

(1) Posterior Data Modeling :

posterior : $\underset{M\times 1}{p(\bar{w}|\bar{t})} \propto \underset{N\times 1}{p(\bar{t}|\bar{w})}\ \underset{M\times 1}{p(\bar{w})}$ (likelihood, prior)

where : Likelihood $p(\bar{t}|\bar{w}) = \prod_n p(t_n|\bar{w}) = \prod_n y_n^{t_n}(1-y_n)^{(1-t_n)}$

Prior $p(\bar{w}) = \underset{M\times 1}{G(\bar{w}; \underset{M\times 1}{\bar{m}_0}, \underset{M\times M}{S_0})}$; $\bar{m}_0 = 0$, $S_0^{-1} = \alpha^{-1}I$

see last page for more discussion

Hence :

$$p(\bar{w}|\bar{t}) \propto p(\bar{t}|\bar{w})\,p(\bar{w}) = \prod_n y_n^{t_n}(1-y_n)^{(1-t_n)}\,G(\bar{w}; \bar{m}_0, S_0)$$

posterior

(2) MAP Error Function:

$$\bar{w}_{MAP} = \underset{\{\bar{w}\}}{\operatorname{argmax}}\ p(\bar{w}|\bar{t}) = \underset{\{\bar{w}\}}{\operatorname{argmin}}\ E(\bar{w})$$

where:

$$E(\bar{w}) = -\log p(\bar{w}|\bar{t}) \quad ; \quad \sigma_n = \sigma(\bar{w}^T\bar{\phi}_n) = y_n$$

$$= -\left[\sum_n t_n \log \sigma_n + (1-t_n)\log(1-\sigma_n)\right] + \frac{1}{2}(\bar{w}-\bar{m}_0)^T S_0^{-1}(\bar{w}-\bar{m}_0) + \text{const}$$

(3) Minimization: $\frac{\partial E(\bar{w})}{\partial \bar{w}} = 0$

$M \times 1$

$$\frac{\partial}{\partial \bar{w}} E(\bar{w}) = -\left[\sum_n t_n \frac{1}{\sigma_n}\sigma_n(1-\sigma_n)\bar{\phi}_n - \sum_n (1-t_n)\frac{\sigma_n}{(1-\sigma_n)}(\sigma_n - 1)\bar{\phi}_n\right] + S_0^{-1}(\bar{w}-\bar{m}_0)$$

($M\times 1$; $\bar{\phi}_n$: $M\times 1$; S_0^{-1}: $M\times M$; $(\bar{w}-\bar{m}_0)$: $M\times 1$)

$$= -\sum_n t_n(1-\sigma_n)\bar{\phi}_n + \sum_n (1-t_n)\sigma_n\bar{\phi}_n + S_0^{-1}(\bar{w}-\bar{m}_0)$$

$$= -\sum_n t_n\bar{\phi}_n + \sum_n t_n\sigma_n\bar{\phi}_n + \sum_n \sigma_n\bar{\phi}_n - \sum_n t_n\sigma_n\bar{\phi}_n + S_0^{-1}(\bar{w}-\bar{m}_0)$$

$$= -\left[\sum_n (t_n\bar{\phi}_n - \sigma_n\bar{\phi}_n)\right] + S_0^{-1}(\bar{w}-\bar{m}_0)$$

$$= S_0^{-1}(\bar{w}-\bar{m}_0) + \sum_n (\sigma_n - t_n)\bar{\phi}_n$$

$$\nabla E(\bar{w}) = \frac{\partial}{\partial \bar{w}} E(\bar{w}) = S_0^{-1}(\bar{w}-\bar{m}_0) + \Phi^T(\bar{\sigma}-\bar{t})$$

($M\times 1$; S_0^{-1}: $M\times M$; $(\bar{w}-\bar{m}_0)$: $M\times 1$; Φ^T: $M\times N$; $(\bar{\sigma}-\bar{t})$: $N\times 1$)

$$H = \nabla\nabla E(\bar{w}) = S_0^{-1} + \sum_n \sigma_n(1-\sigma_n)\bar{\phi}_n\bar{\phi}_n^T$$

(H: $M\times M$; $\nabla\nabla E(\bar{w})$: $M\times M$; S_0^{-1}: $M\times M$; $\bar{\phi}_n$: $M\times 1$; $\bar{\phi}_n^T$: $1\times M$)

$$H = S_0^{-1} + \Phi R \Phi^T$$

($M\times M$ = $M\times M$ + $M\times N$ $N\times N$ $N\times M$)

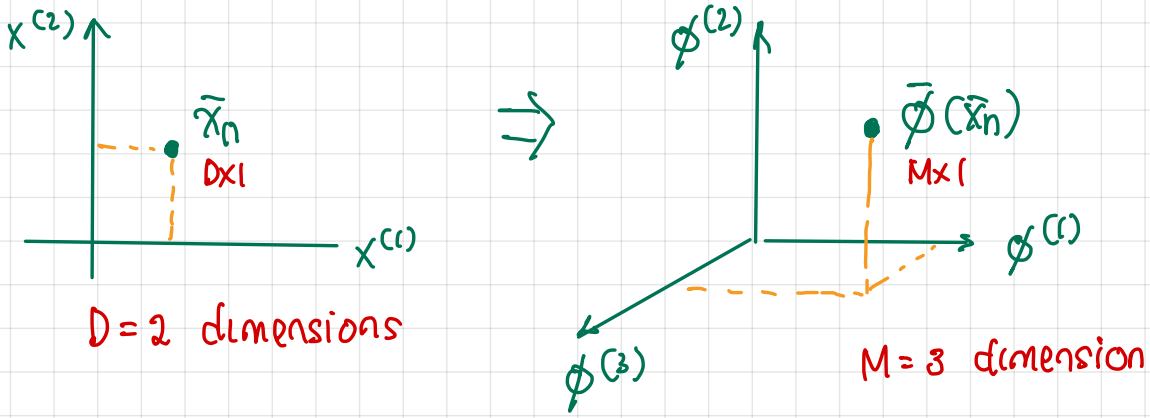
$$R_{nn} = \sigma_n(1-\sigma_n)$$

Newton's method: $\bar{w}^{new} = \bar{w}^{old} - H^{-1}\nabla E(\bar{w})$

Notes :

1. Q: Why do we model $p(\bar{w}) = G(\bar{w}; 0, \alpha \mathbb{I})$? What is the meaning of this prior ?
   ($\bar{w}$: M×1, $\mathbb{I}$: M×M)

A: In regression, the model complexity M determines the complexity of the fitting line. In classification, M determines the dimensionality of the transform space. The classification line / plane is always linear.



Hence, large M implies a space with high dimensionality, where we can always separate the positive from negative classes. This implies overfitting.

↓↓

To reduce the overfitting problem, we encourage some of w's to be zero (= reducing the dimensions)

FULL BAYESIAN CLASSIFICATION

Training stage: Input: $\{\bar{x}_n, t_n\}_{n=1}^{N}$; $\bar{t} = \begin{bmatrix} t_1 \\ t_2 \\ \vdots \\ t_N \end{bmatrix}$

Output: $\bar{w}$

(i) Full Bayesian Data Modeling:

Posteriori: $p(\bar{w}|\bar{t}) = \dfrac{p(\bar{t}|\bar{w})\, p(\bar{w})}{p(\bar{t})}$ (likelihood, prior)

Likelihood: $p(\bar{t}|\bar{w}) = \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}$

Prior: $p(\bar{w}) = G(\bar{w}; \bar{m}_0, S_0)$

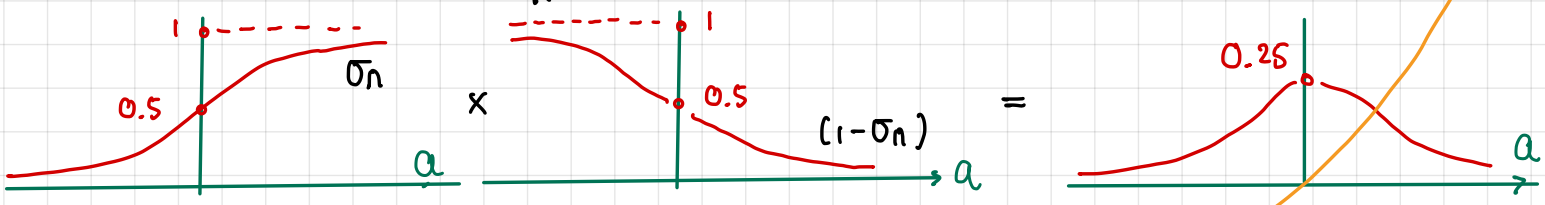
Hence:

$$p(\bar{w}|\bar{t}) = \frac{\prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}\, G(\bar{w}; \bar{m}_0, S_0)}{\int \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}\, G(\bar{w}; \bar{m}_0, S_0)\, d\bar{w}}$$

Problems:

1. In regression, we can solve the full Bayesian equation using Case 2. Case 2 requires both likelihood & prior to be Gaussians. Yet, in the above equation, the likelihood is NOT Gaussian.

2. Likelihood: $p(\bar{t}|\bar{w}) = \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}$ is not Gaussian:



3. $f(\bar{w}) = \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)}\, G(\bar{w}; \bar{m}_0, S_0)$

looks like a Gaussian

[•] Laplace Approximation

Goal: To approximate $f(\bar{w})$ with a Gaussian function.

$$f(\bar{w}) = \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)} \; G(\bar{w}; \bar{m}_0, S_0)$$

$$\approx \underbrace{q(\bar{w}) = G(\bar{w}; \bar{m}_N, S_N)}_{\text{to find}}$$

Concept:

$$p(x) = \frac{1}{Z} f(x) \quad ; \quad Z = \int f(x)\, dx$$

Z: unknown normalization factor

$$q(x) \approx p(x)$$

To find: $q(x)$, a Gaussian appoximation

Step 1:

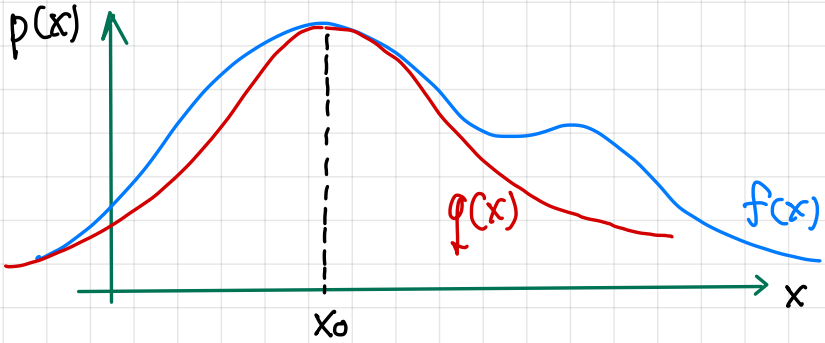
Mean of $q(x)$ = mode of $f(x)$

x_0 = mode of $f(x)$

thus: $\left.\frac{d}{dx} f(x)\right|_{x=x_0} = 0$

Step 2:

Covariance of $q(x)$ = covariance computed from $f(x)$



(mode of $f(x)$)

$f(x) \rightarrow p(x)$ has exactly the same shape as $f(x)$

Step #2 : To find the approximated covariance from $f(x)$

Let : $g(x) = \log f(x)$

Taylor expansion :

$$g(x) = g(x_0) + (x-x_0)\frac{d}{dx}g(x)\Big|_{x=x_0} + \frac{(x-x_0)^2}{2}\frac{d^2}{dx^2}g(x)\Big|_{x=x_0}$$

(Idea : to use this Taylor expansion to find the covariance from $f(x)$)

where : x_0 = the mode of $f(x)$: $\frac{d}{dx}f(x)\Big|_{x=x_0} = 0$

Implying : $\frac{d}{dx}g(x) = \frac{d}{dx}\log f(x) = \frac{1}{f(x)}\frac{d}{dx}f(x) = 0$

Hence :

$$g(x) = g(x_0) + \underbrace{(x-x_0)\frac{d}{dx}g(x)\Big|_{x=x_0}}_{=0} + \frac{(x-x_0)^2}{2}\frac{d^2}{dx^2}g(x)\Big|_{x=x_0}$$

$$= g(x_0) + \frac{(x-x_0)^2}{2}\frac{d^2}{dx^2}g(x)\Big|_{x=x_0}$$

$$\log f(x) = \log f(x_0) + \frac{(x-x_0)^2}{2}\frac{d^2}{dx^2}\log f(x)\Big|_{x=x_0}$$

Let $$A = -\frac{d^2}{dx^2}\log f(x)\Big|_{x=x_0}$$

Then : $$\log f(x) = \log f(x_0) - \frac{A}{2}(x-x_0)^2$$

$$f(x) = f(x_0)\exp\left(-\frac{A}{2}(x-x_0)^2\right)$$

Therefore : The precision from $f(x) = \frac{1}{\sigma^2} = \beta = A$

σ^2 = standard deviation

If the precision of $q(x)$ = A, then:

$$q(x) = \left(\frac{A}{2\pi}\right)^{1/2} \exp\left(-\frac{A}{2}(x-x_0)^2\right) \quad ; \quad A = -\frac{d^2}{dx^2}\log f(x)$$

$$\Downarrow$$

for vector $\bar{x}$: $\quad p(\bar{x}) = \frac{1}{Z} f(\bar{x})$

$$\underset{D\times 1}{q(\bar{x})} = \frac{|\underset{D\times D}{A}|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\bar{x}-\bar{x}_0)^T A (\bar{x}-\bar{x}_0)\right)$$

$$= G(\underset{D\times 1}{\bar{x}}; \underset{D\times 1}{\bar{x}_0}, \underset{D\times D}{A^{-1}})$$

Back to the problem we have:

$$f(\bar{w}) = \prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)} \, G(\bar{w}; \bar{m}_0, S_0)$$

$$\approx \boxed{q(\bar{w}) = G(\bar{w}; \bar{m}_N, S_N)} \quad \text{to find}$$

$$\Downarrow$$

Q: What is $\bar{m}_N$? What is S_N?

Recall from Lecture Note 18:

$$E(\bar{w}) = -\log\left(p(\bar{t}|\bar{w})\, p(\bar{w})\right)$$

$$= -\log\left[\prod_n \sigma_n^{t_n} (1-\sigma_n)^{(1-t_n)} \, G(\bar{w}; \bar{m}_0, S_0)\right]$$

Hence: $\quad E(\bar{w}) = -\log f(\bar{w})$

If $E(\bar{w}) = -\log f(\bar{w})$:

Step 1 :

$\bar{m}_N$ (the mode of $f(\bar{w})$) $= \bar{w}_{MAP}$

Because $\bar{w}_{MAP}$ means the $\bar{w}$ that indicates the highest point of $E(\bar{w})$ or $f(\bar{w})$.

Step 2 :

Covariance of $q(\bar{w}) = S_N$, where $S_N = A^{-1}$

$$A = -\nabla\nabla \log f(\bar{w})$$

$$= \nabla\nabla(-\log f(\bar{w})) = \nabla\nabla E(\bar{w})\Big|_{\bar{w}=\bar{w}_{MAP}}$$

$$= H = S_0^{-1} + \underset{M\times M}{\Phi^T} \underset{M\times N}{R} \underset{N\times M}{\Phi}$$

(see lecture note 18)

or: Textbook Eq. (4.143) page 218.

Therefore :

$$f(\bar{w}) = p(\bar{t}|\bar{w})\, p(\bar{w}) \approx q(\bar{w}) = G(\bar{w}; \bar{w}_{MAP}, S_N)$$

$$= \frac{|H|^{1/2}}{(2\pi)^{D/2}} \exp\left(-\frac{1}{2}(\bar{w}-\bar{w}_{MAP})^T \underset{D\times D}{H^{-1}} (\bar{w}-\bar{w}_{MAP})\right)$$

Finally :

$$p(\bar{w}|\bar{t}) = \frac{p(\bar{t}|\bar{w})\, p(\bar{w})}{p(\bar{t})} = q(\bar{w})$$

Note: $\int q(\bar{w})\, d\bar{w} = 1$